UNIVERZITET U BANJOJ LUCI
PRIRODNO-MATEMATIČKI FAKULTET
MATEMATIKA I INFORMATIKA, INFORMATIKA
UVOD U ISTRAŽIVANJE PODATAKA

ANALIZA MUZIČKIH TRENDOVA IZ PERIODA 2000-2019. NA MUZIČKOJ PLATFORMI *SPOTIFY*

Seminarski rad

Student:
Sara Talić (15/19)

Predmetni profesor:
doc. dr Milana Grbić

Banja Luka, jun 2024.

S A D R Ž A J

1. UVOD

Spotify je švedska platforma za strimovanje muzike i pružanje medijskih usluga, osnovana 2006. godine, a javnosti je postala dostupna 2008. godine. Ova platforma omogućava korisnicima pristup širokom katalogu pjesama, albuma i podkasta putem interneta. Korisnici mogu besplatno slušati muziku sa reklamama ili platiti pretplatu za *premium* verziju koja uključuje dodatne funkcije poput slušanja bez reklama i slušanja bez povezivanja na internet [1, 2].

Spotify se ističe sa 615 miliona aktivnih korisnika na mjesečnom nivou, preko 100 miliona pjesama, 6 miliona podkasta i 350 000 audio knjiga. Platforma je dostupna u 180 različitih regiona [1].

Ovaj rad ima za cilj da istraži i analizira muzičke trendove iz perioda od 2000. do 2019. godine na *Spotify* platformi, uključujući evoluciju žanrova, popularnost izvođača, kao i preferencije slušalaca u pogledu specifičnih karakteristika muzike kao što su tempo, energičnost ili tekstualna tema. Kroz analizu podataka, ovaj rad će predstaviti ključne trendove koji su obilježili muzičku scenu tokom protekle dvije decenije.

2. ANALIZA I PRIPREMA PODATAKA

Podaci korišćeni za istraživanje u ovom radu, preuzeti su sa *Kaggle* sajta, 13. juna 2024. godine. Ovaj skup podataka sadrži statistike audio-zapisa najboljih 2000 pjesama na *Spotify* platformi iz perioda od 2000. do 2019. godine [3].

2.1 Opis originalnih podataka

Skup podataka sastoji se od 2000 redova i 18 kolona.

Kolone su sljedeće:

artist - Ime izvođača pjesme,

song - Naziv pjesme,

duration_ms - Trajanje pjesme u milisekundama,

explicit - Da li pjesma sadrži eksplicitne tekstove ili sadržaj koji se smatra neprikladnim za djecu (False/True),

year - Godina izdavanja pjesme,

popularity - Ocjena popularnosti pjesme, pri čemu veća vrijednost označava veću popularnost,

danceability - Opisuje koliko je pjesma pogodna za ples, vrijednost 0.0 je najmanje pogodna za ples, a 1.0 najviše,

energy – Mjeri se intenzitet i aktivnost pjesme, vrijednosti od 0.0 do 1.0,

key - Tonalitet pjesme, izražen u numeričkom obliku koji mapira tonove u određene ključeve, npr. C=0, D=2, ako nije identifikovan ključ vrijednost je -1,

loudness - Ukupna glasnoća pjesme izražena u decibelima (dB), vrijednosti su između -60 i 0 dB,

mode - Označava modus (major ili minor) pjesme, vrijednost 1 za major ili 0 za minor,

speechiness - Prisustvo govora u pjesmi, vrijednost bliža 1 ukazuje na veći udio govornog sadržaja,

acousticness - Mjera koliko je pjesma akustična, vrijednost bliža 1 označava veću akustičnost,

instrumentalness - Da li pjesma sadrži vokale ili je instrumentalna, vrijednost bliža 1 ukazuje na veću vjerovatnoću da je pjesma instrumentalna,

liveness - Prisustvo publike u snimku, vrijednost bliža 1 ukazuje na to da je pjesma snimljena uz prisustvo publike,

valence - Mjeri koliko pozitivno ili negativno pjesma zvuči, vrijednost bliža 1 ukazuje na pozitivniji ton,

tempo - Ukupna brzina pjesme izražena u udarima po minuti (BPM) i

genre - Žanr pjesme [3].

2.2 Analiza podataka i pretprocesiranje za dalju analizu

Analiza podataka započeta je provjerom *Null* vrijednosti, ovaj skup podataka ih nema. Potom je uočeno prisustvo 59 dupliranih instanci, te su iste izbrisane.

Iako se u naslovu skupa podataka navodi opseg 2000-2019. izvršićemo provjeru da li su sve godine upravo iz tog opsega. Provjerom je utvrđeno da postoje numere iz 1998, 1999. i 2020. godine (Slika 2.1). Kako bismo ostali dosljedni naslovu našeg rada, sve numere koje nisu izdate između 2000. i 2019. godine će biti izbrisane. Ovim je izbrisano još 42 instance.

	Year	Number of songs
0	1999	38
1	2020	3
2	1998	1

Slika 2.1. Broj numera koje nisu u opsegu 2000-2019.

S obzirom na to da istražujemo muzičke trendove, te ispitujemo najpopularnije pjesme, iz skupa podataka ćemo da izbrišemo pjesme čija je ocjena popularnosti jednaka nuli. Takvih pjesama je 126.

Pošto su sva imena kolona napisana malim slovima, izvršili smo izmjenu tako da svaki naziv počinje velikim slovom (Slika 2.2).

```
#Izmjena naziva kolona, capslk

data.rename(columns={'artist':'Artist', 'song':'Song', 'duration_ms':'Duration_MS', 'explicit':'Explicit',
                     'year':'Year','popularity':'Popularity','danceability':'Danceability',
                     'energy':'Energy', 'key':'Key', 'loudness':'Loudness', 'mode':'Mode',
                     'speechiness':'Speechiness', 'acousticness':'Acousticness',
                     'instrumentalness':'Instrumentalness', 'liveness':'Liveness',
                     'valence':'Valence', 'tempo':'Tempo','genre':'Genre'},inplace=True)
```

Slika 2.2. Izmjena naziva kolona

Počevši od 2000 instanci, nakon prethodnih koraka, analizu nastavljamo sa 1773 instanci.

2.2.1 Istraživačka analiza podataka

Istraživačka analiza podataka (engl. *Exploratory data analysis*) je ključni korak u procesu istraživanja i razumijevanja podataka. Ovo uključuje detaljno istraživanje skupa podataka kako bismo identifikovali obrasce, tendencije i bitne karakteristike. Često se koriste metode vizualizacije kako bismo olakšali razumijevanje [4].

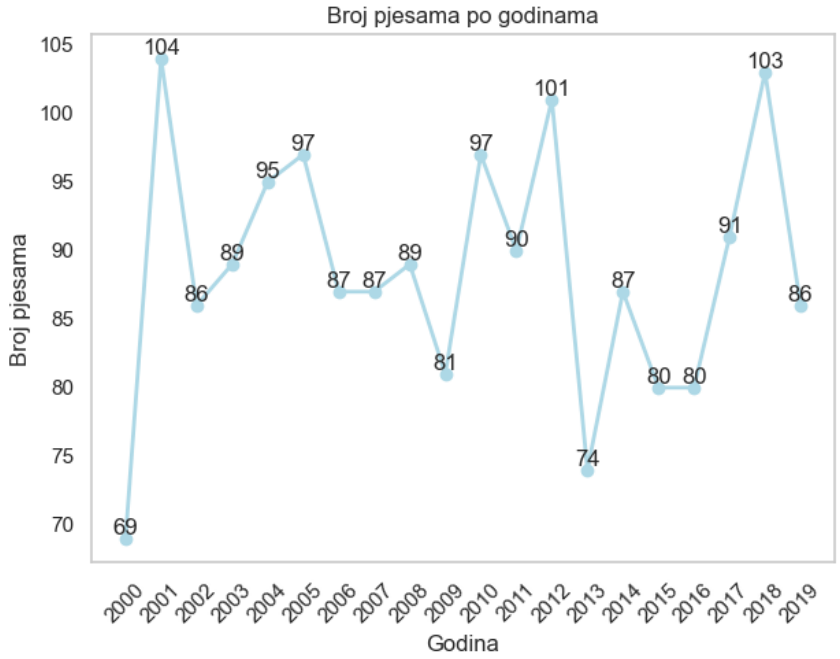
U nastavku su prikazana neka istraživanja do kojih smo došli.

Sva analiza je rađena u *Jupyter Notebook-u*, pri čemu su korišćene sledeće biblioteke: *pandas, numpy, matplotlib.pyplot, seaborn* i *plotly.express.*

Na Slici 2.3 prikazani su rezultati istraživanja broja pjesama po svakoj godini od 2000. do 2019. Da bismo lakše uočili razlike, na Grafiku 2.1 vizuelno su predstavljeni dobijeni rezultati.

	Year	Number of songs
0	2000	69
1	2001	104
2	2002	86
3	2003	89
4	2004	95
5	2005	97
6	2006	87
7	2007	87
8	2008	89
9	2009	81
10	2010	97
11	2011	90
12	2012	101
13	2013	74
14	2014	87
15	2015	80
16	2016	80
17	2017	91
18	2018	103
19	2019	86

Slika 2.3. Rezultati istraživanja broja pjesama po godinama



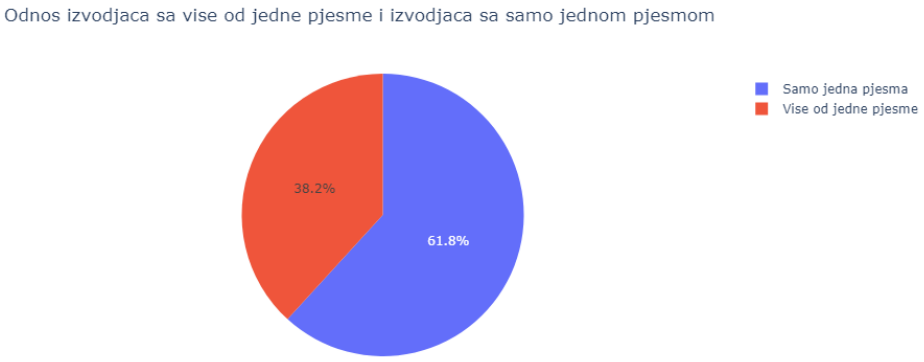
Grafik 2.1. Broj pjesama po godinama

U skupu podataka nalazi se 788 različitih izvođača. Na Slici 2.4 prikazano je deset izvođača sa najviše numera.

	Artist	Number of songs
595	Rihanna	23
222	Eminem	21
117	Calvin Harris	20
204	Drake	20
182	David Guetta	18
110	Britney Spears	17
680	Taylor Swift	16
82	Beyoncé	16
371	Katy Perry	15
367	Kanye West	15

Slika 2.4. Deset izvođača sa najviše numera

Ovi rezultati podstakli su sljedeći grafik na kojem je vršeno poređenje izvođača sa samo jednom numerom i izvođača sa više od jedne numere. Rezultati su predstavljeni uz pomoć *pie chart*-a i prikazani su na Grafiku 2.2.



Grafik 2.2. Odnos izvođača sa više od jedne pjesme i izvođača sa samo jednom pjesmom

Sljedeće istraživanje koje je izvršeno je deset najpopularnih pjesama. Na Slici 2.5 prikazani su rezultati istraživanja, prikazani su nazivi pjesama, izvođači, godine izdavanje pjesama i ocjene popularnosti (mjere se od 0 do 100).

	Artist	Song	Year	Popularity
1229	The Neighbourhood	Sweater Weather	2013	89
1220	Tom Odell	Another Love	2013	88
192	Eminem	Without Me	2002	87
1475	WILLOW	Wait a Minute!	2015	86
6	Eminem	The Real Slim Shady	2000	86
1641	Billie Eilish	lovely (with Khalid)	2018	86
1814	Post Malone	Circles	2019	85
1151	Bruno Mars	Locked out of Heaven	2012	85
1551	Ed Sheeran	Perfect	2017	85
1393	Avicii	The Nights	2014	85

Slika 2.5. Deset najpopularnijih pjesama iz perioda 2000-2019. na *Spotify-u*

Ovi rezultati su bili inspiracija sljedećeg istraživanja. Zanimalo nas je koji su to izvođači imali najpopularnije numere u ove dvije decenije. Na Slici 2.6 prikazani su rezultati, kako bismo dobili rezultate sumirali smo ocjenu popularnosti svih numera izvođača. Međutim, ako se sjetimo

rezultata sa Slike 2.4, saznali smo da ova tri izvođača nemaju isti broj numera u ovom skupu podataka. S tim na umu, računati sumu vrijednosti popularnosti i nije najbolji izbor, te smo se odlučili da posmatramo prosječnu popularnost za ova tri izvođača, kako bismo vidjeli razliku, ti rezultati su prikazani na Slici 2.7.

	Artist	Total Popularity
1	Rihanna	1662
2	Eminem	1519
3	Drake	1424

Slika 2.6. Prva tri izvođača sa najpopularnijim numerama

	Artist	Number of Songs	Mean Popularity
1	Eminem	21	72.333333
2	Rihanna	23	72.260870
3	Drake	20	71.200000

Slika 2.7. Poredak izvođača sa Slike 2.6 računajući prosječnu ocjenu popularnosti

Izvođači koji su ostvarili najveću popularnost sa svojim numerama prikazani su na Slici 2.8. Primjetno je da je većina izvođača sa samo jednom numerom na ovoj listi.

Artist	Mean Popularity	Number of Songs
Tom Odell	88.0	1
The Neighbourhood	87.0	2
WILLOW	86.0	1
Lewis Capaldi	84.0	1
Gesaffelstein	84.0	1
Tame Impala	83.0	1
MKTO	82.0	1
Alec Benjamin	82.0	1
girl in red	82.0	1
Foster The People	82.0	1

Slika 2.8. Deset izvođača čije su numere ostvarile najveću popularnost

Zašto je ovakva analiza podataka korisna, možemo da vidimo na sljedećem primjeru. Sljedeće istraživanje koje je vršeno jeste broj pjesama po žanrovima. Važno je naglasiti da neke pjesme pripadaju više od jednom žanru, te su takve pjesme uračunate u sve žanrove koji ih karakterišu. Na Slici 2.9 pod *a)* prikazani su rezultati istraživanja, zahvaljujući ovim rezultatim primijetili smo

da postoje pjesme čija je vrijednost kolone žanr jednaka ‘set()’. Žanr pjesama kod kojih se ovo pojavilo zamijenili smo sa vrijednošću *‘Other’*.

Genre	
pop	1442
hip hop	707
R&B	406
Dance/Electronic	332
rock	208
metal	61
latin	56
set()	21
country	17
Folk/Acoustic	17
Name: count, dtype: int64	

a)

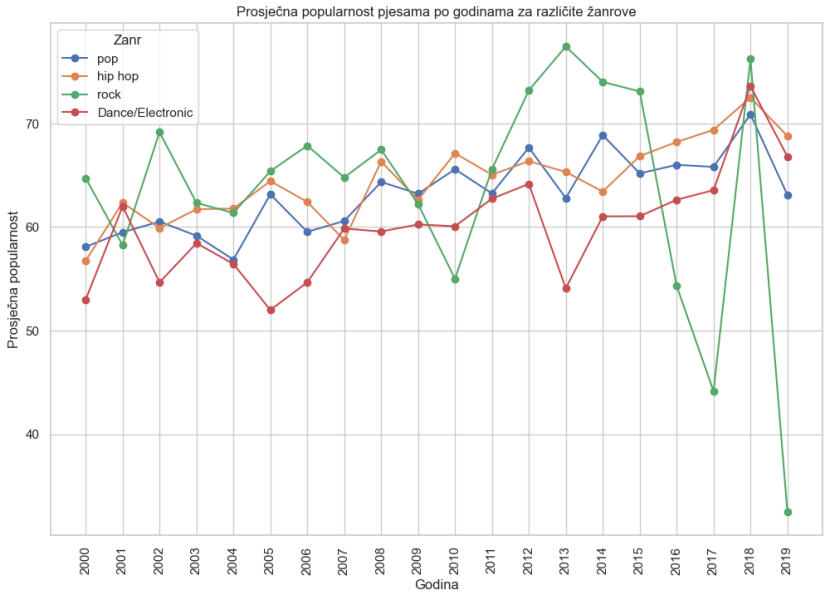
Genre	
pop	1442
hip hop	707
R&B	406
Dance/Electronic	332
rock	208
metal	61
latin	56
Other	21
country	17
Folk/Acoustic	17
Name: count, dtype: int64	

b)

Slika 2.9. Broj pjesama po žanrovima, a) prvi rezultati, primjećena je neadekvatna vrijednost ‘set()’, b) rezultati nakon što smo vrijednosti ‘set()’ zamijenili sa ‘Other’

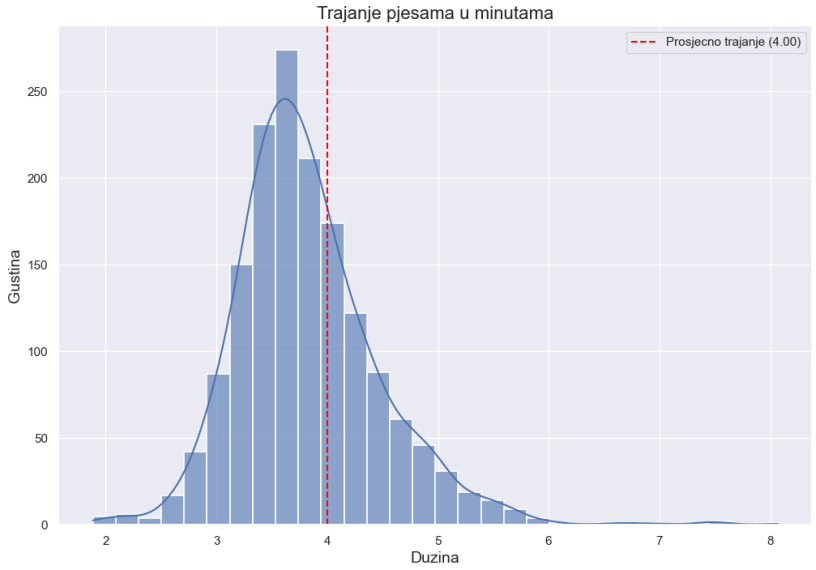
Nakon prethodnog istraživanja provjerene su i ostale kolone, kako bi se uvidjelo da li su vrijednosti u adekvatnim opsezima. Nismo naišli na probleme u ostalim kolonama.

Istražili smo i popularnost pjesama po žanrovima za svaku godinu, ovi rezultati prikazani su na Grafiku 2.3.

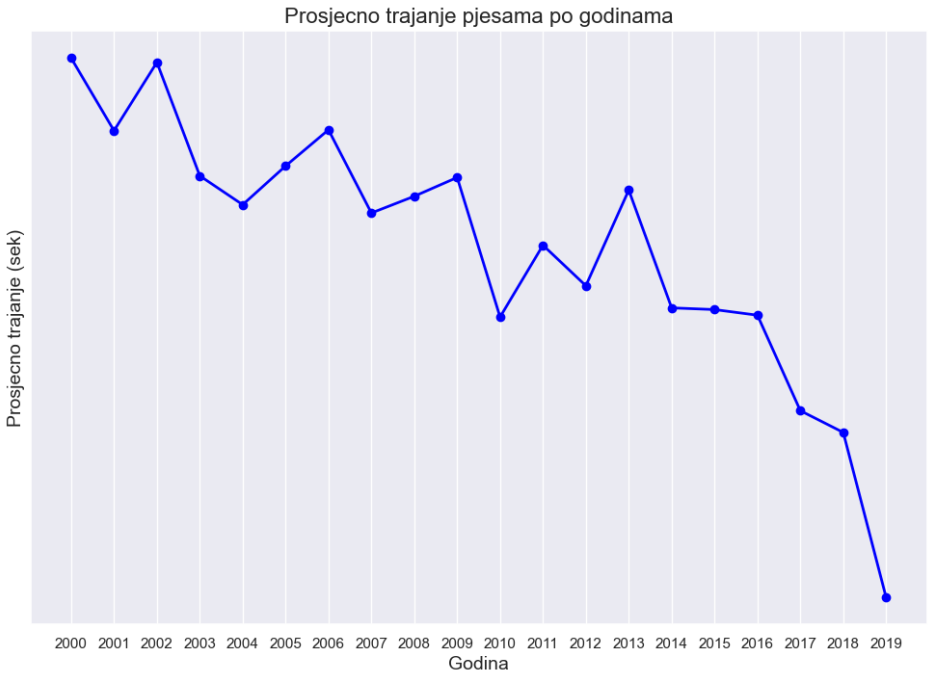


Grafik 2.3. Popularnost žanrova tokom godina

Sljedeća istraživanja vezana su za trajanje numera. Na Grafiku 2.4 prikazana je raspodjela trajanja pjesama. Naredni grafik, Grafik 2.5, predstavlja prosječno trajanje pjesama po godinama, gdje je primjetan pad trajanja pjesama, odnosno pjesme su sve kraće.

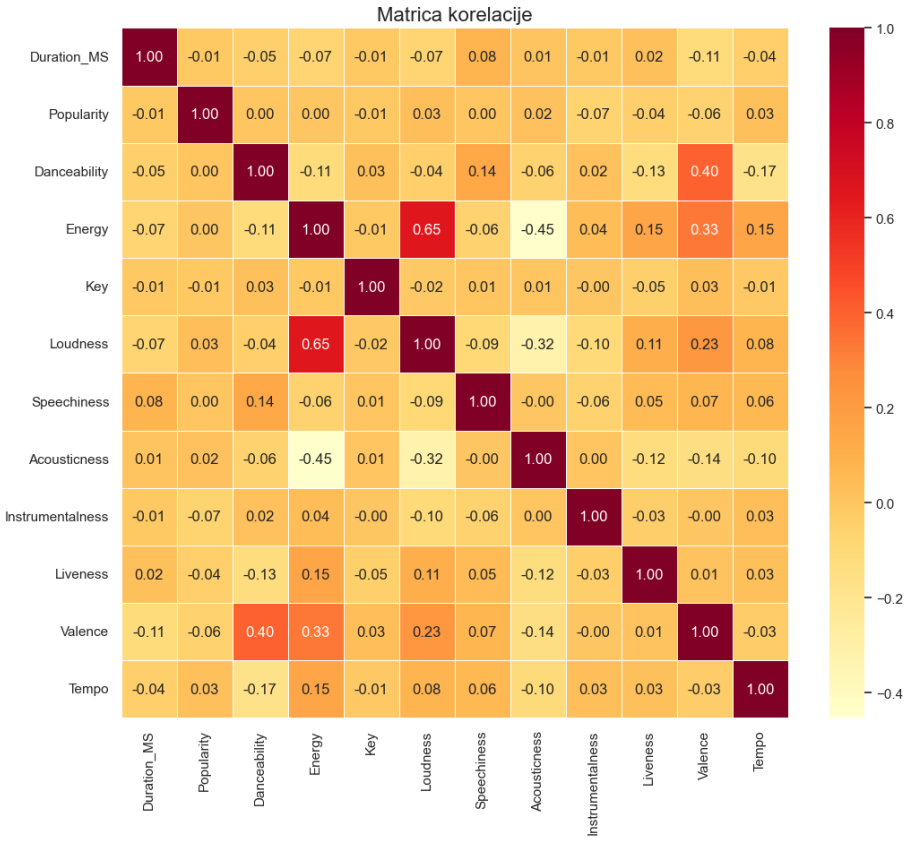


Grafik 2.4. Prosječno trajanje pjesama u minutama



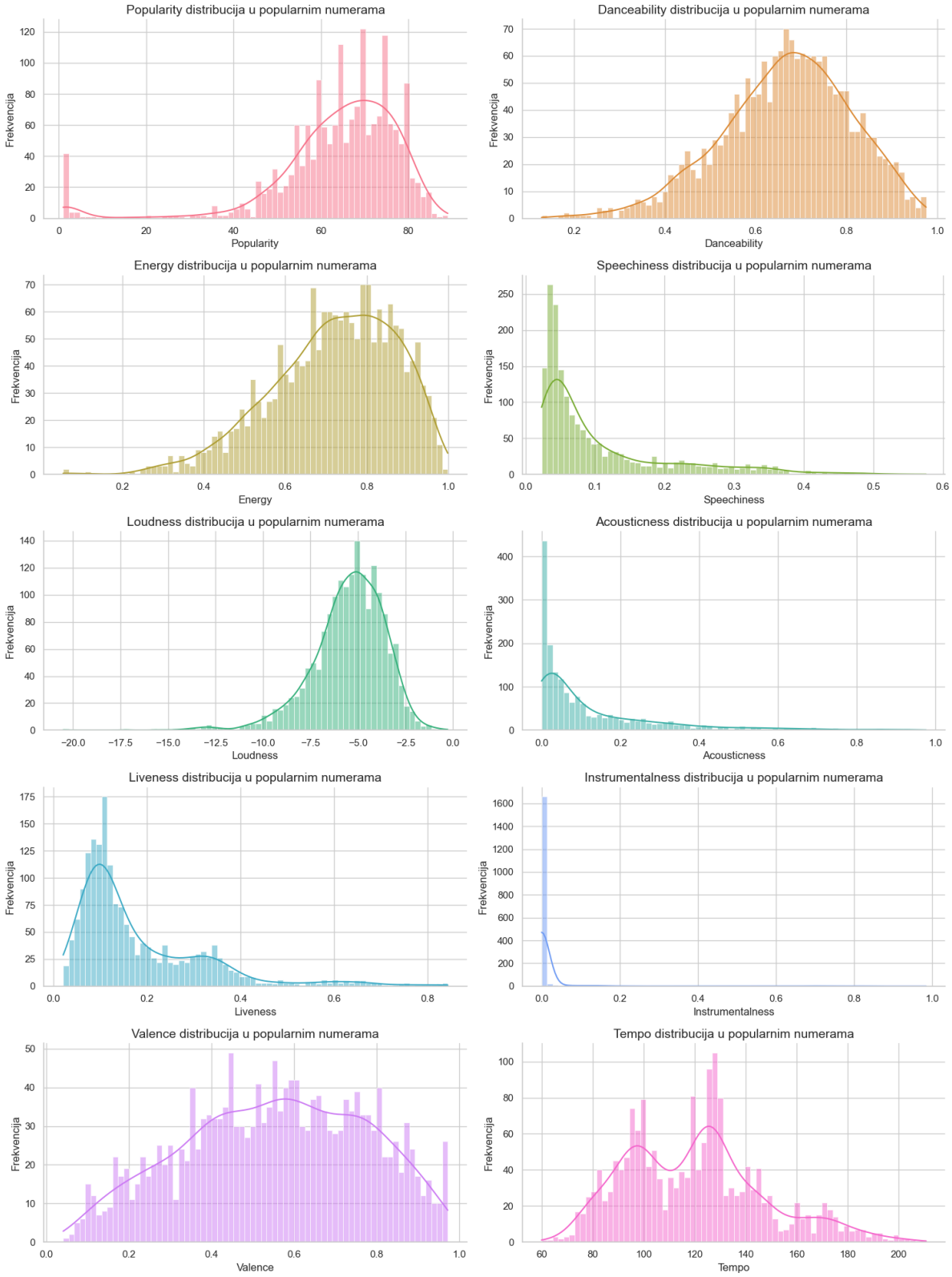
Grafik 2.5. Prosječno trajanje pjesama po godinama

Većina kolona su numeričke kolone vezane za audio-karakteristika same pjesme. Takve kolone posmatrali smo putem matrice korelacije, kako bismo uvidjeli potencijalne povezanosti kolona, rezultati su prikazani na Grafiku 2.6.



Grafik 2.6. Matrica korelacije numeričkih kolona

Za kraj, istraživali smo distribuciju popularnosti i audio-karakteristika u numerama našeg skupa podataka. U obzir su uzete sljedeće kolone: *'Popularity', 'Danceability', 'Energy', 'Speechiness', 'Loudness', 'Acousticness', 'Liveness', 'Instrumentalness', 'Valence'* i *'Tempo'*. Na Grafiku 2.7 prikazani su rezultati istraživanja.



Grafik 2.7. Distribucija audio-karakteristika

3. METODE

Kako bismo poboljšali našu analizu koristićemo algoritme klasterovanja.

K-means algoritam

K-means algoritam je jednostavan algoritam zasnovan na prototipovima. U ovom algoritmu prototip se definiše u vidu centroida. Centroid je srednja vrijednost grupe tačaka. *K-means* funkioniše tako što prvo odabere K početnih centroida, te se potom svaka tačka dodjeljuje najbližem centroidu čime se formiraju klasteri [5].

Broj početnih centroida u našem radu odabran je na dva načina:

1. Koeficijent sjenke (engl. *Silhouette Coefficient*)

   Ovaj metod sastoji se od nekoliko koraka:

   - Za neku tačku *i*, izračuna se njegova udaljenost od svih drugih tačaka u njegovom klasteru, ta vrijednost je označena sa *a,*
   - Za tačku *i* i bilo koji klaster koji ne sadrži tu tačku izračuna se prosječna udaljenost tačke *i* do svih tačaka odabranog klastera, ta vrijednost je označena sa *b,*
   - Koeficijent sjenke za tačku *i* računa se formulom $s = \frac{b-a}{\max(a,b)}$ [5].

   Vrijednost koeficijenta sjenke se nalazi u opsegu od -1 do 1. Što je vrijednost *a* bliža 0, to je vrijednost koeficijenta sjenke bliža 1 [5].

2. Metod lakta (engl. *Elbow Method*)

   Ovaj metod podrazumijeva iteriranje kroz različite vrijednosti K od 1 do n. Za svaku vrijednost se računa zbir kvadrata rastojanja tačaka do njihovog najbližeg centra klastera. Grafik izgledom podsjeća na oblik lakta, te biramo onu vrijednost gdje grafik počinje ličiti na pravu liniju [6].

Inicijalizaciju željenog broja centroida možemo da izvršimo nasumično ili uz pomoć *k-means++* algoritma. *K-means++* omogućava pametniju inicijalizaciju centroida, što dovodi do boljeg kvaliteta samih klastera [7].

Koraci dodjeljivinja tačaka centroidima ponavljaju se dok ne dođe do centroida koji se ne mijenjaju [5].

Hijerarhijsko klasterovanje

Pored *K-means* algoritma, hijerarhijsko klasterovanje je takođe značajna tehnika. Dva načina za generisanje ovog klasterovanja su:

- Algoritmi sakupljanja (engl. *Agglomerative*) – na početku je svaka tačka zaseban klaster, te se u svakom koraku spaja najbliži par klastera,
- Algoritmi razdvajanja (engl. *Divisive*) – suprotno algoritmu sakupljanja, odnosno počinjemo sa jednim klasterom kojeg u svakom korako dijelimo dok ne dobijemo pojedinačne tačke.

Ovo klasterovanje najčešće se prikazuje u vidu grafika koji se naziva dendogram [5].

Postoji nekoliko načina za definisanje blizine klastera:

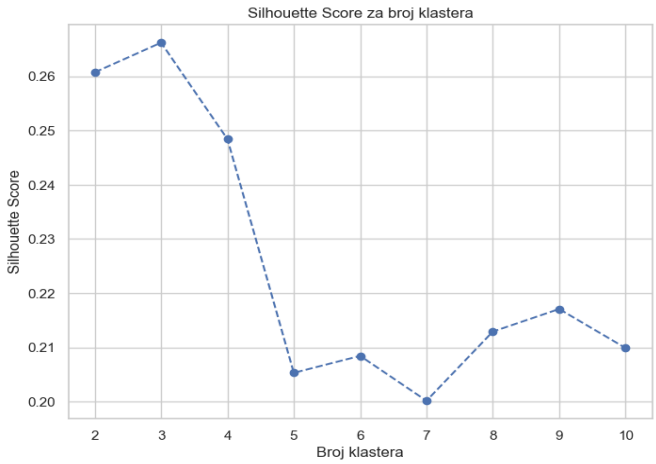
- MIN ili *Single linkage* – blizina klastera jeste blizina između dvije najbliže tačke koje pripadaju različitim klasterama,
- MAX ili *Complete linkage* – blizina klastera je blizina između dvije najudaljenije tačke koje pripadaju različitim klasterima,
- *Average linkage* – blizina klastera je prosječna blizina svih parova tačaka dva različita klastera,
- *Ward* metod – ovaj metod teži ka tome da minimizuje zbir kvadratnih udaljenosti tačaka od njihovih centroida [5].

4. REZULTATI

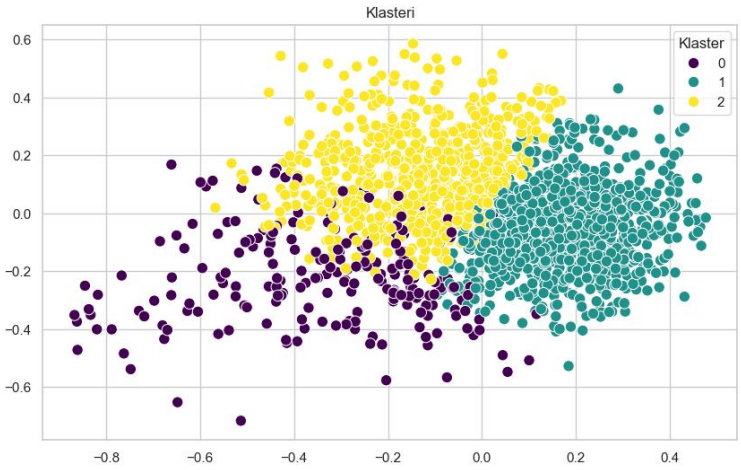
Sva klasterovanja su vršena na osnovu sljedećih audio-karakteristika *Energy, Acousticness, Valence, Instrumentalness* i *Tempo.* Odabrane su karakteristike koje međusobno nisu povezane, te na osnovu njih želimo da pokušamo da grupišemo pjesme i uvidimo sličnost pjesama iz istih grupa.

4.1 *K-means* algoritam. Klasterovanje uz pomoć *Silhouette Coeficient*

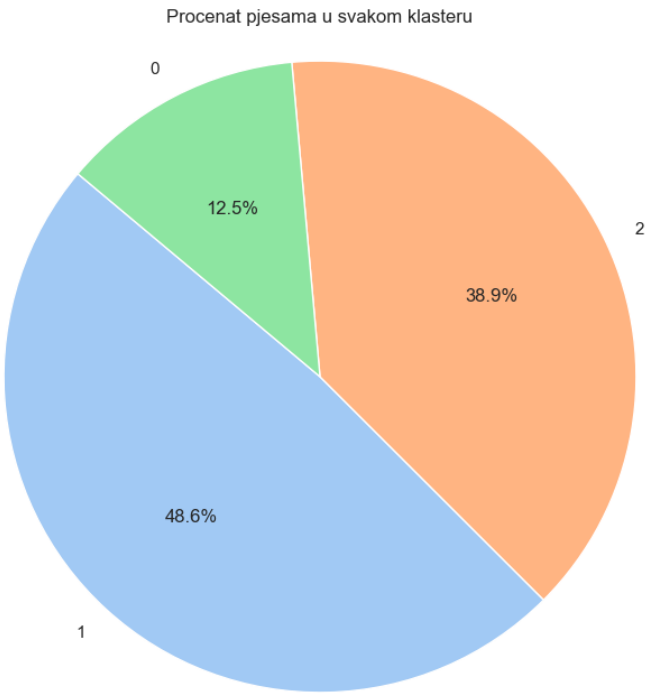
Koeficijent sjenke najviše je iznosio 0.266 za broj klastera tri (Grafik 4.1). Ovo nije zadovoljavajući rezultat, ali s ozbirom na naše podatke iz skupa, razmotrićemo klasterovanje sa tri klastera. Za inicijalizaciju centroida nasumičan izbor i *k-means*++ algoritam pokazali su skoro identične rezultate, te ćemo mi prikazati rezultate dobijeni korišćenjem *k-means*++ algoritma.



Grafik 4.1. Koeficijent sjenke, *K-means* algoritam



Grafik 4.2. Klasterovanje *K-means* na osnovu dobijenog optimalnog broja klastera uz pomoć koeficijenta sjenke

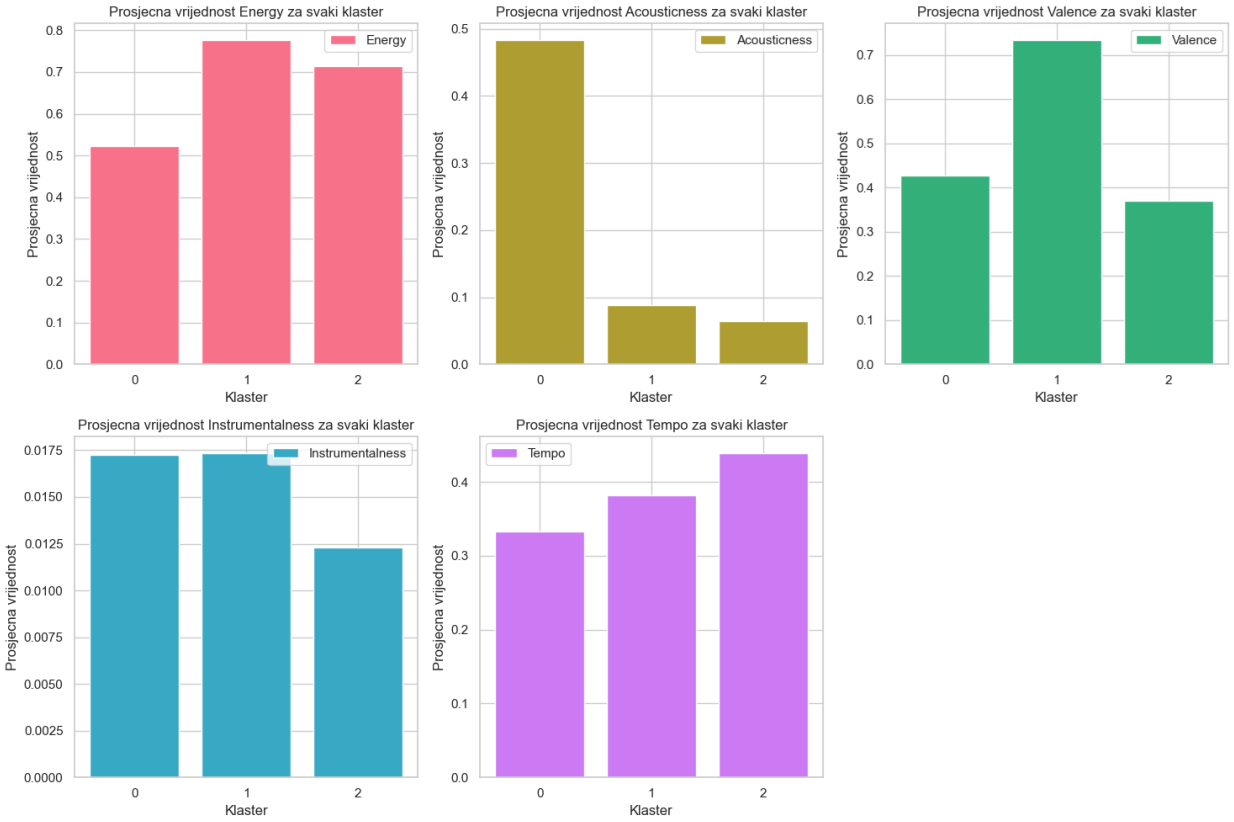


Grafik 4.3. Raspodjela pjesama po klasterima, *K-means* algoritam uz pomoć koeficijenta sjenke

Cluster	Energy	Acousticness	Valence	Instrumentalness	Tempo
0	0.522973	0.483525	0.427900	0.017271	0.332381
1	0.777028	0.088475	0.734358	0.017353	0.381912
2	0.712970	0.064514	0.370563	0.012317	0.439097

Slika 4.1. Prosječne vrijednosti audio-karakteristika, *K-means* algoritam uz koeficijent sjenke

Prethodni grafici nisu izgledali kao najzadovoljavajući. Međutim, kada pogledamo Grafik 4.4 na kom su prikazane prosječne vrijednosti audio-karakteristika, uočljivo je da kreirani klasteri imaju smisla, te da svaki klaster predstavlja slične pjesme, odnosno pjesme istog ili sličnog žanra.



Grafik 4.4. Prosječne audio-karakteristike po klasterima, koeficijent sjenke

Analiza rezultata

Pjesme koje pripadaju klasteru 0 imaju najviši nivo akustičnosti, instrumentalnosti i najmanje su energične. Takođe, njihov tempo je najslabiji, i manje su pozitivne. Možemo zaključiti da su se u ovom klasteru pronašle lagane pjesme, uglavnom balade. To potvrđuje i činjenica da su neke od pjesama iz ovog klastera upravo takve numere:

Another Love - Tom Odell

lovely - Billie Eilish, Khalid

Perfect - Ed Sheeran

Do I Wanna Know? - Artic Monkeys.

Pjesme koje pripadaju klasteru 1 jesu najenergičnije i najpozitivnije. Odlikuju se instrumentalnošću, manjom akustičnošću i bržim tempom. Potrebno je istaći da su ovakve numere najbrojnije, u procentu od 48.6%, što znači da su najpopularnije numere upravo ovih karakteristika. Ovaj klaster obuhvatio je numere kao što su:

Without Me - Eminem

The Real Slim Shady - Eminem

Locked out of Heaven - Bruno Mars

The Nights – Avicii.

Klaster 2 karakterišu najmanje akustične, energične pjesme, najbržeg tempa. Takođe, numeru su umjereno pozitivne i najmanje instrumentalne. Pop numere odlikuju se, uglavnom, ovakvim karakteristima, međutim u ovom klasteru je i dosta numera koje nisu odgovarele ni prvom ni drugom klasteru. Neke od numera iz ovog klastera su:

Sweater Weather - The Neighbourhood

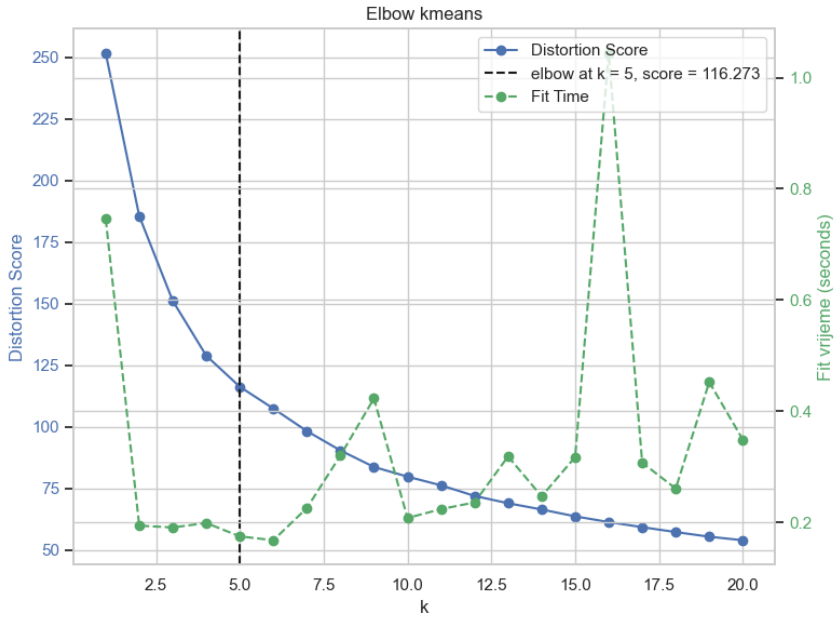
The Hills - The Weeknd

One Dance - Drake

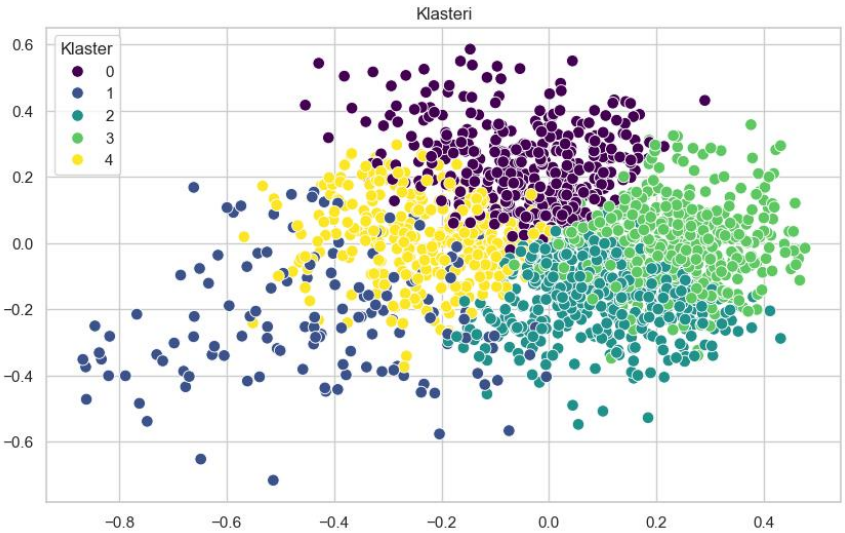
No Lie - Sean Paul.

4.2 *K-means* algoritam. Klasterovanje uz pomoć *Elbow* metoda

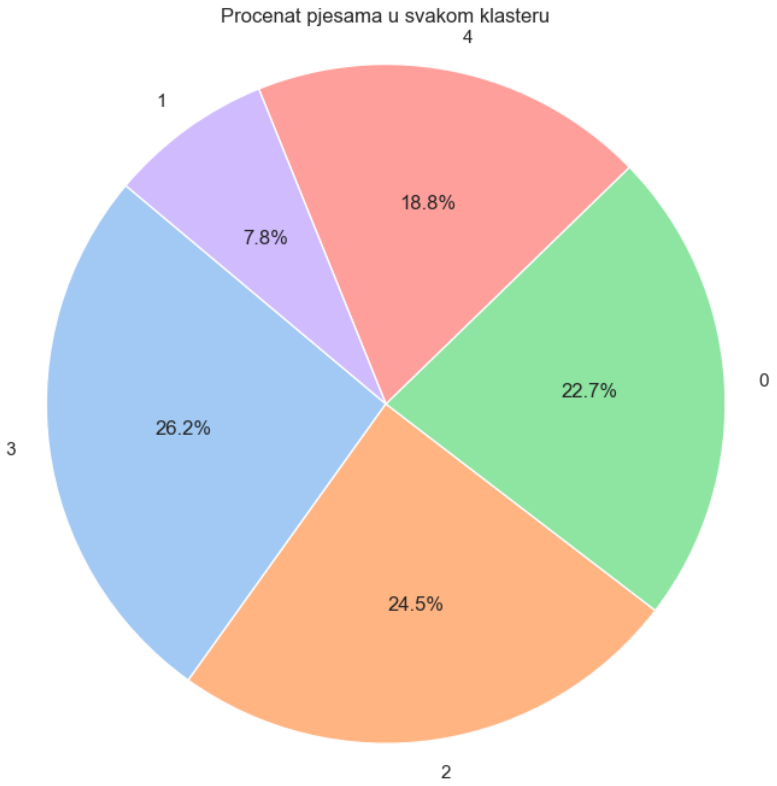
Kako bismo dodatno analizirali naš problem, korišćen je i metod lakta, odnosno *Elbow* metod, za određivanje optimalnog broja klastera. *Elbow* metodom izračunat je optimalan broj klastera pet (Grafik 4.5). Kao i u prethodnom primjeru i ovdje su nasumičan izbor i *k-means*++ algoritam za inicijalizaciju centroida pokazali skoro identične rezultate, te će biti prikazani rezultati *k-means*++ algoritma.



Grafik 4.5. *Elbow* metod, *K-means* algoritam



Grafik 4.6. Klasterovanje *K-means* na osnovu dobijenog optimalnog broja klastera uz pomoć *Elbow* metoda

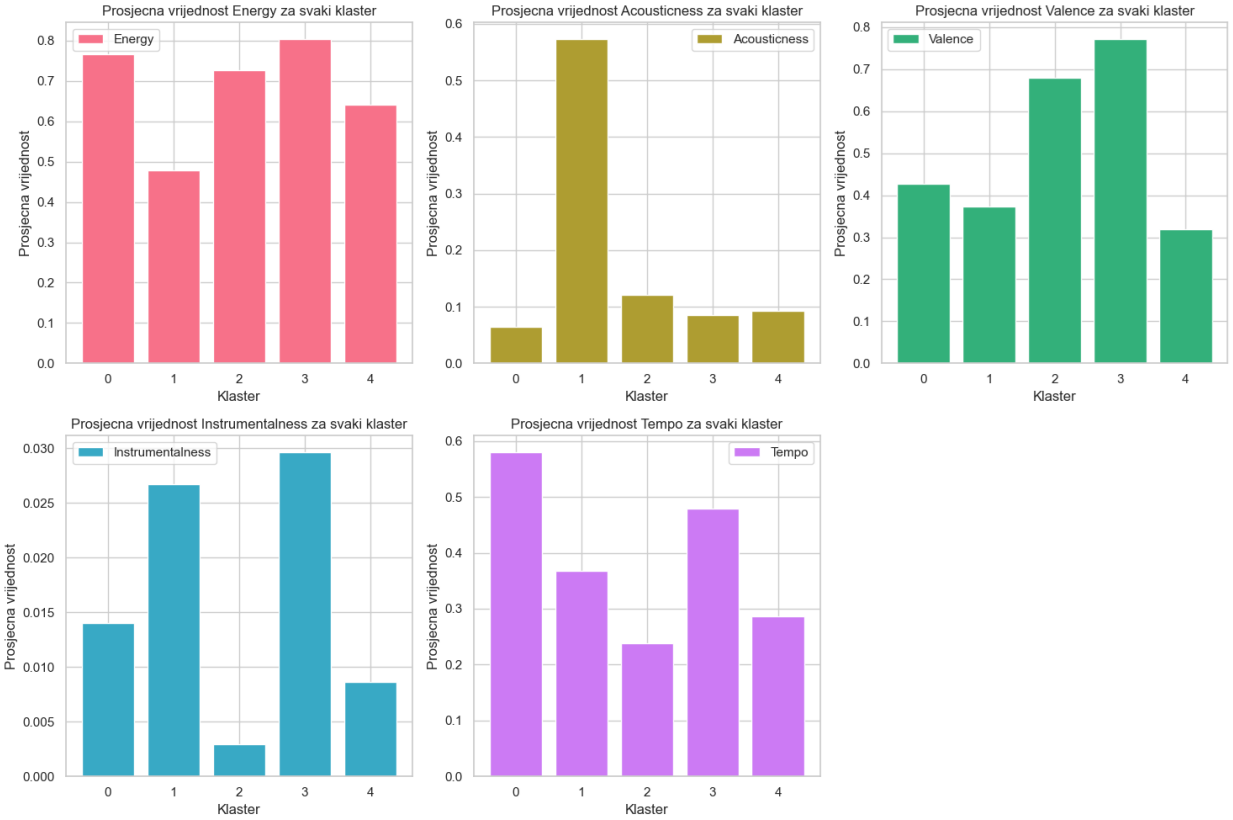


Grafik 4.7. Raspodjela pjesama po klasterima, *K-means* algoritam uz pomoć *Elbow* metoda

ClusterE	Energy	Acousticness	Valence	Instrumentalness	Tempo
0	0.766189	0.064067	0.427290	0.014041	0.580467
1	0.478174	0.573630	0.372167	0.026715	0.367509
2	0.726539	0.120401	0.679028	0.002928	0.237536
3	0.804662	0.084984	0.772432	0.029670	0.479182
4	0.640246	0.092669	0.318153	0.008609	0.286427

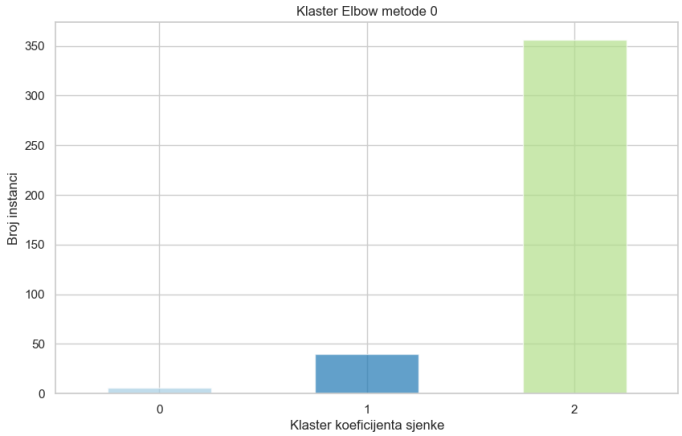
Slika 4.2. Prosječne vrijednosti audio-karakteristika, *K-means* algoritam uz *Elbow* metod

Ni ovim klasterovanjem prethodni grafici nisu izgledali obećavajuće, međutim kada pogledamo Grafik 4.8 ponovo uočavamo smislene sličnosti numera u klasterima. Za razliku od prethodnog klasterovanja, ovdje su zahvaljujući većem broju klastera numere više odvojene po karakteristikama.

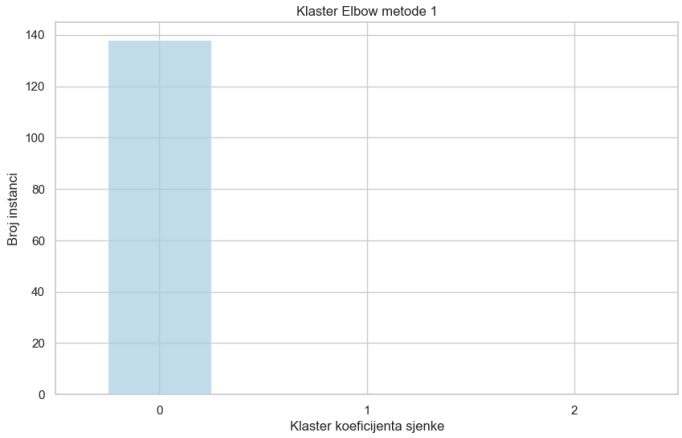


Grafik 4.8. Prosječne audio-karakteristike po klasterima, *Elbow* metod

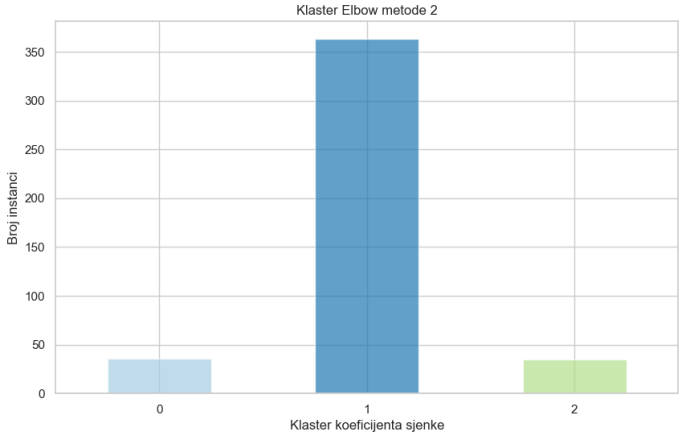
Analiza rezultata



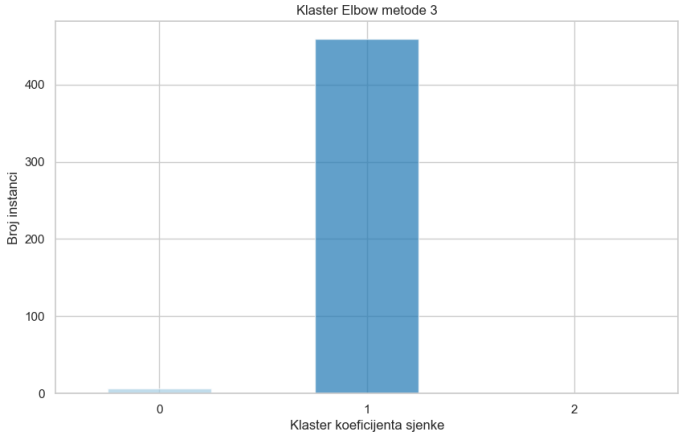
Grafik 4.9. Raspodjela klastera dobijenih uz pomoć klastera sjenke u klasteru 0 *Elbow* metoda



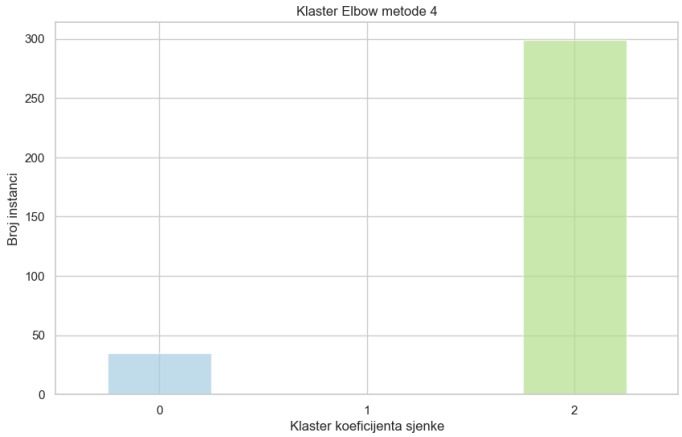
Grafik 4.10. Raspodjela klastera dobijenih uz pomoć klastera sjenke u klasteru 1 *Elbow* metoda



Grafik 4.11. Raspodjela klastera dobijenih uz pomoć klastera sjenke u klasteru 2 *Elbow* metoda



Grafik 4.12. Raspodjela klastera dobijenih uz pomoć klastera sjenke u klasteru 3 *Elbow* metoda



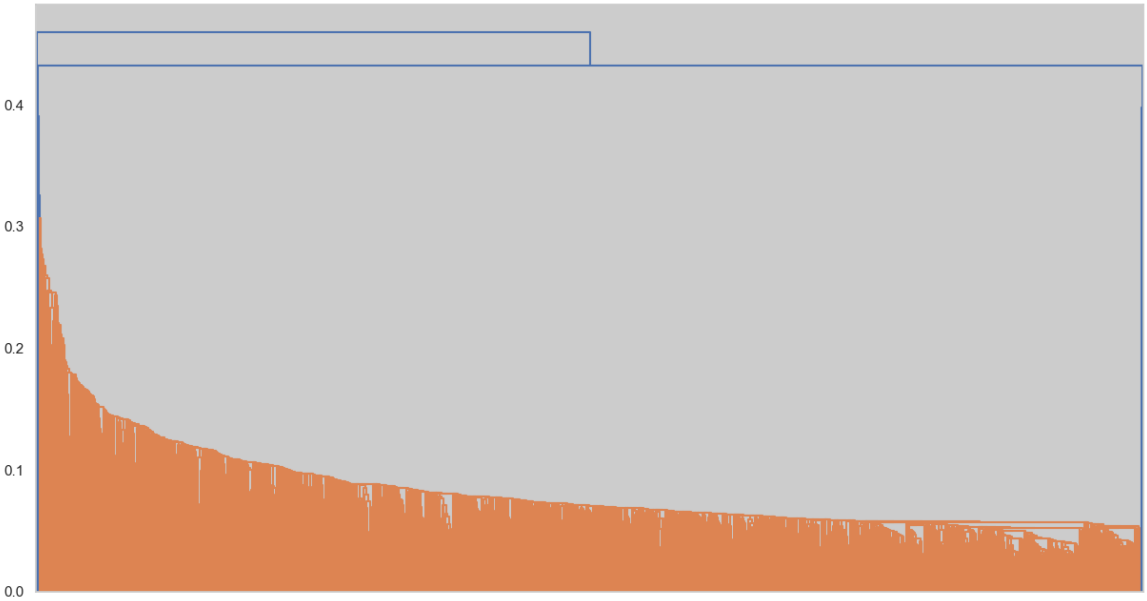
Grafik 4.13. Raspodjela klastera dobijenih uz pomoć klastera sjenke u klasteru 4 *Elbow* metoda

Za razliku od koeficijenta sjenke, *Elbow* metodom dobijen je optimalan broj klastera pet. U prethodnim graficima (Grafik 4.9-Grafik 4.13) ispitali smo raspodjelu klastera koeficijenta sjenke u novodobijenim klasterima.

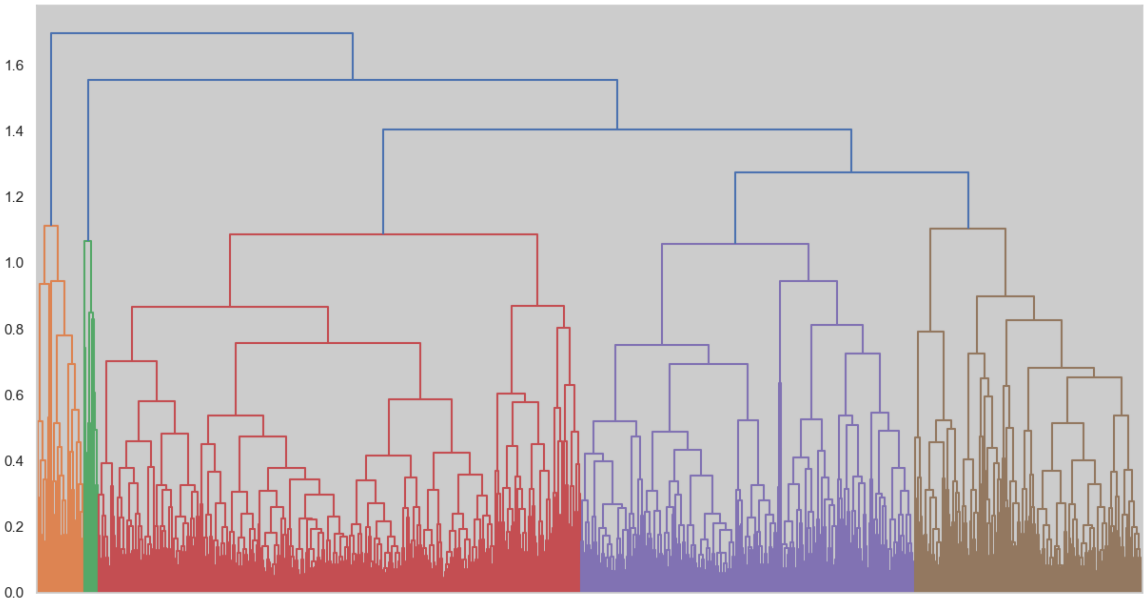
Klasteri 1 i 2 koeficijenata sjenke su podijeljeni među više klastera *Elbow* metoda, čime smo dobili preciznije klasifikovanje pjesama. Klaster 0 koeficijenta sjenke odgovara Klasteru 1 *Elbow* metoda, a tu su najakustičnije pjesme, sa najmanje energije, kojih ujedno ima i najmanje.

4.3 Hijerarhijsko klasterovanje

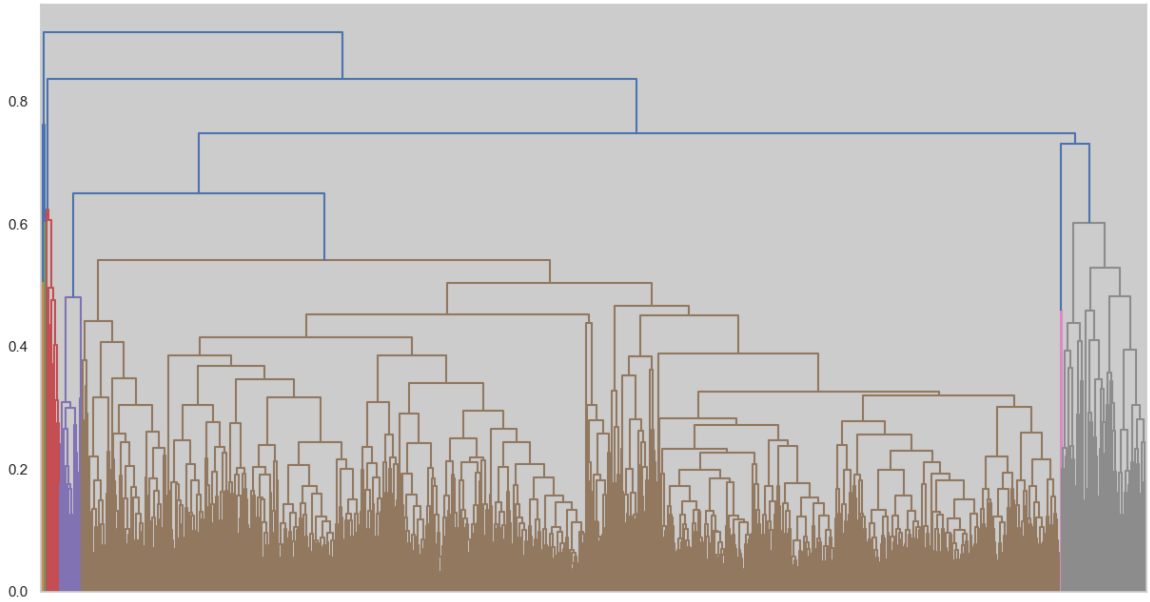
Na sljedećim graficima (Grafik 4.14-Grafik 4.17) prikazani su dendogrami dobijeni korišćenjem različitih definisanja blizina klastera.



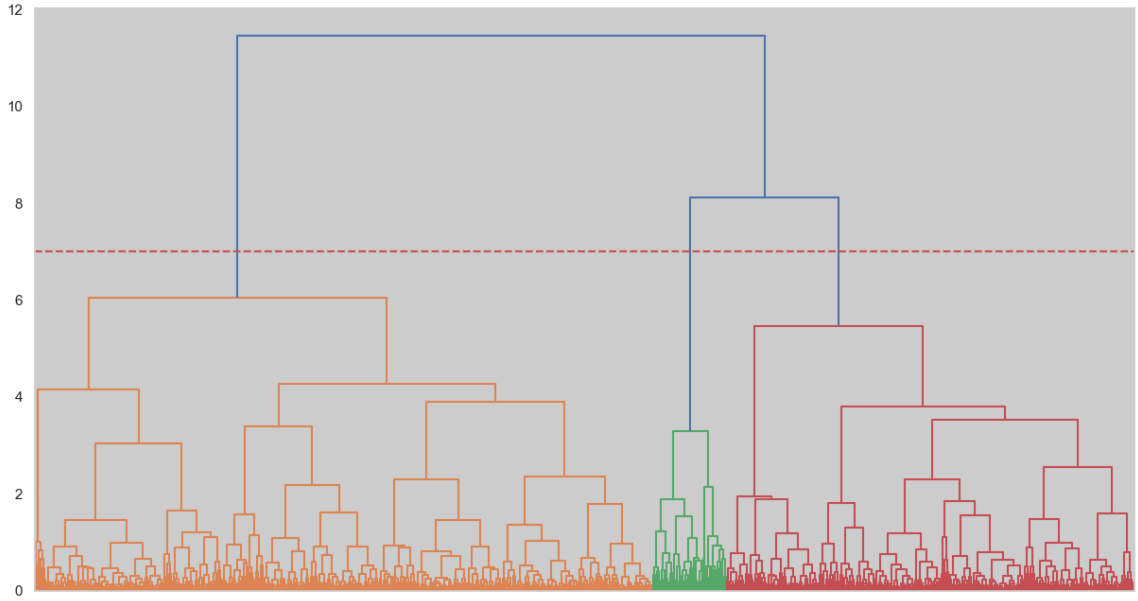
Grafik 4.14. Dendogram hijerarhijskog klasterovanja korišćenjem *Single linkage*



Grafik 4.15. Dendogram hijerarhijskog klasterovanja korišćenjem *Complete linkage*

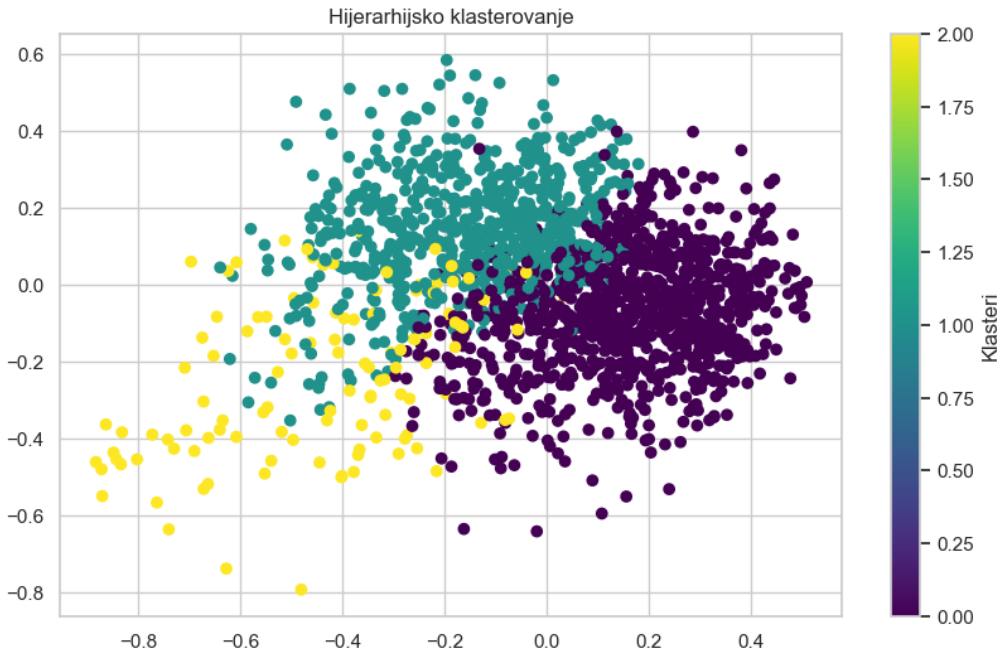


Grafik 4.16. Dendogram hijerarhijskog klasterovanja korišćenjem *Average linkage*

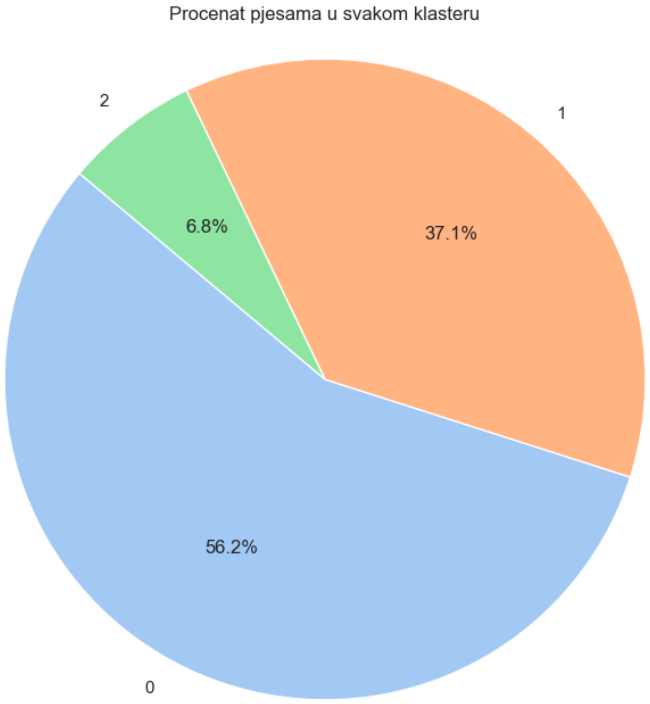


Grafik 4.17. Dendogram hijerarhijskog klasterovanja korišćenjem *Ward* metoda

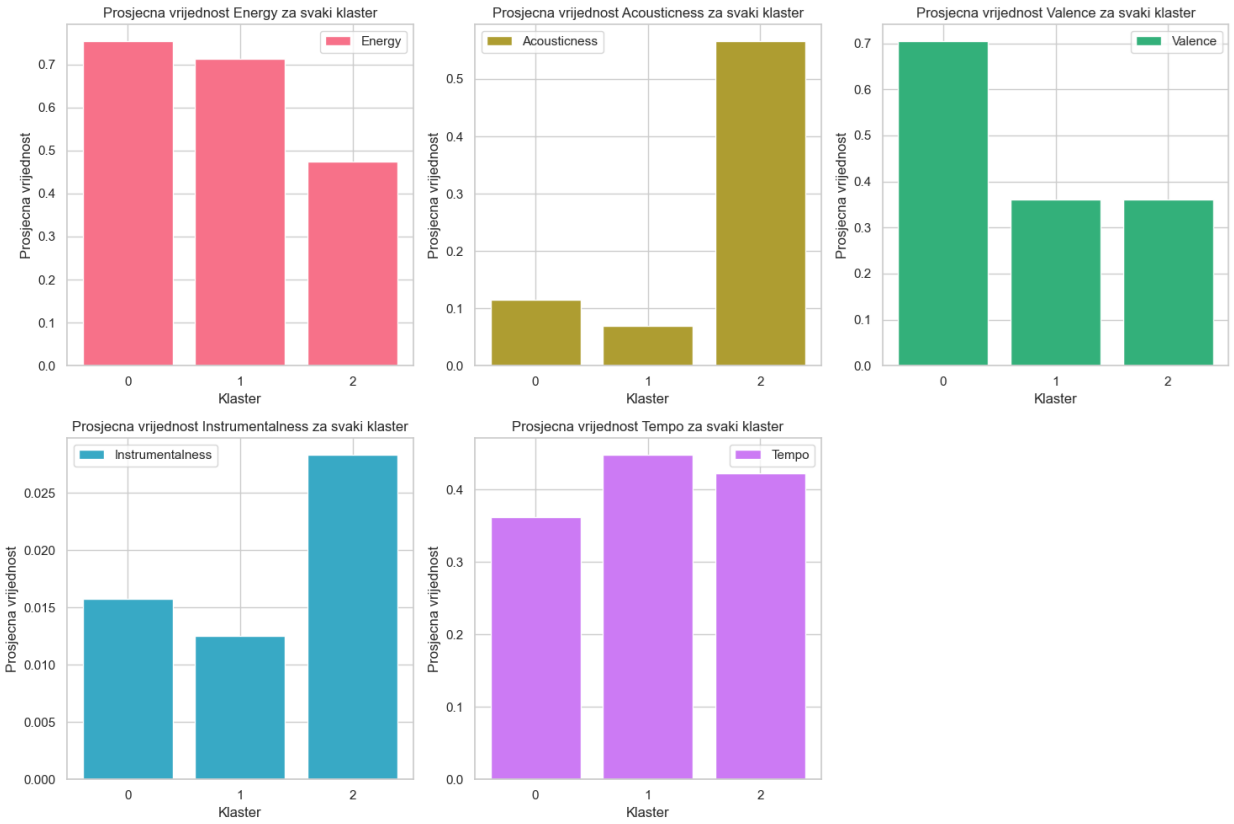
Nakon pregleda dobijenih dendograma, najbolji rezultat je dobijen *ward* metodom, gdje su pjesme grupisane u tri klastera. U nastavku ćemo prikazati dobijene klastere ovom metodom.



Grafik 4.18. Klasteri, hijerarhijsko klasterovanje



Grafik 4.19. Raspodjela pjesama po klasterima, hijerarhijsko klasterovanje



Grafik 4.20. Prosječne audio-karakteristike po klasterima, hijerarhijsko klasterovanje

4.4 Poređenje rezultata *K-means* algoritma i hijerarhijskog klasterovanja

Analiziraćemo uporedno rezultate dobijene korišćenjem *K-means* algoritma i hijerarhijskog klasterovanja. Pošto je broj klastera u hijerarhijskom klasterovanju bio tri, poredićemo ove klastere sa klasterima *K-means* algoritma uz korišćenje koeficijenta sjenke (Potpoglavlje 4.1).

Klasteru 0 hijerarhijskog klasterovanja odgovara Klaster 1 *K-means* algoritma. Klaster 1 *K-means* algoritma zauzeo je 48.6% raspodjele, dok je Klaster 0 hijerarhijskog zauzeo 56.2%. Ovim klasterima pripadaju najenergičnije i najpozitivnije pjesme. Odlikuju se instrumentalnošću i manjom akustičnošću

Tabela 4.1. Poređenje Klastera 1 *K-means* algoritma i Klastera 0 hijerarhijskog klasterovanja

	K-means	**Hijerarhijsko**
Energy	0.777	0.754
Acousticness	0.088	0.115
Valence	0.734	0.705
Instrumentalness	0.017	0.016
Tempo	0.381	0.361

Klasteru 1 hijerarhijskog klasterovanja odgovara Klaster 2 *K-means* algoritma. Klaster 2 *K-means* algoritma zauzeo je 38.9% raspodjele, dok je Klaster 1 hijerarhijskog 37.1%. Numere ovih klastera su slabo akustične, energične pjesme, najbržeg tempa. Takođe, numeru su umjereno pozitivne i najmanje instrumentalne.

Tabela 4.2. Poređenje Klastera 2 *K-means* algoritma i Klastera 1 hijerarhijskog klasterovanja

	K-means	**Hijerarhijsko**
Energy	0.713	0.714
Acousticness	0.065	0.069
Valence	0.371	0.362
Instrumentalness	0.013	0.013
Tempo	0.439	0.439

Klasteru 2 hijerarhijskog klasterovanja odgovara Klaster 0 *K-means* algoritma. Klaster 0 *K-means* algoritma zauzeo je 12.5% raspodjele, dok je Klaster 2 hijerarhijskog 6.8%. Ovim klasterima pripadaju balade ili slične numere, kojih je u obje raspodjele najmanje.

Tabela 4.3. Poređenje Klastera 0 *K-means* algoritma i Klastera 2 hijerarhijskog klasterovanja

	K-means	**Hijerarhijsko**
Energy	0.523	0.475
Acousticness	0.484	0.567
Valence	0.427	0.361
Instrumentalness	0.017	0.028
Tempo	0.332	0.422

5. ZAKLJUČAK

Analiza muzičkih trendova na platformi *Spotify* u periodu od 2000. do 2019. godine omogućila nam je da bolje razumijemo karakteristike popularnih numera. Uspjeli smo da uvidimo kakvim karakteristikama se odlikuju najveći broj popularnih numera iz ovog perioda.

Detaljna analiza karakteristika pjesama, kao što su tempo, energija, plesnost i drugi atributi, omogućila nam je bolje razumijevanje osnovnih komponenti koje definišu različite muzičke stilove.

Korišćenjem *K-means* algoritma i koeficijenta sjenke podijelili smo popularne numere u tri grupe, te uvidjeli da su najbrojnije numere najenergičnije, najpozitivnije i bržeg tempa. Zahvaljujući metodu lakta grupisali smo numere u pet klastera, što je omogućilo detaljniju podjelu pjesama na osnovu zajedničkih karakteristika. Međutim, i veći broj klastera pokazao je da isti tip pjesama čini najbrojniju grupu popularnih pjesama.

Kako bismo obogatili našu analizu koristilo smo i hijerarhijsko klasterovanje koje nam je pružilo dodatni uvid u strukturu podataka. Uz pomoć dendograma, dodatno smo potvrdili rezultate dobijene *K-means* algoritmom. Hijerarhijskim klasterovanjem smo takođe dobili tri klastera, ali drugačije raspoređenih. Bez obzira na drugačiju raspodjelu, i dalje su najmnogobrojnije numere upravo najenergičnije, brzog tempa i najpozitivnije.

Rezultati ove analize mogu imati značajne implikacije za muzičku industriju. Identifikacija klastera pjesama može pomoći muzičkim platformama poput *Spotify*-a da unaprijede svoje algoritme za preporučivanje muzike i kreiranje plejlista. Takođe, ovakvom analizom bi se moglo potencijalno izvršiti predviđanje da li će neka numera biti popularna ili ne.

Nažalost, naš skup podataka obuhvata pjesme samo do 2019. godine. Počevši od naredne godine, 2020. godine, došlo je do ekspanzije društvene mreže *TikTok*, zahvaljući kojoj su se mnoge numere probile na prva mjesta svjetskih top lista. Interesantno bi bilo u analizu uključiti pjesme iz perioda od 2020. do danas, te ispitati uticaj ove društvene mreže na postavljanje novih muzičkih trendova i potencijalno mijenjanje dosadašnjih karakteristika najpopularnijih pjesama.

LITERATURA

[1] About Spotify, Spotify, [Na mreži], Dostupno: https://newsroom.spotify.com/company-info/ [Pristupljeno 13. 6. 2024].
[2] Spotify, Wikipedia, [Na mreži], Dostupno: https://en.wikipedia.org/wiki/Spotify [Pristupljeno 13. 6. 2024].
[3] Top Hits Spotify from 2000-2019, Kaggle, [Na mreži], Dostupno: https://www.kaggle.com/datasets/paradisejoy/top-hits-spotify-from-20002019/data [Pristupljeno 13. 6. 2024].
[4] What is EDA?, IBM, [Na mreži], Dostupno: https://www.ibm.com/topics/exploratory-data-analysis [Pristupljeno 13. 6. 2024].
[5] Pang.N Ing Tan, Michael Steinbach, Vipin Kumar, Introduction to Data Mining, 2006.
[6] Elbow Method for optimal value of k in KMeans, GeaksforGeeks, [Na mreži], Dostupno: https://www.geeksforgeeks.org/elbow-method-for-optimal-value-of-k-in-kmeans/ [Pristupljeno 18. 6. 2024].
[7] ML | K-means++ Algorithm, GeaksforGeeks, [Na mreži], Dostupno https://www.geeksforgeeks.org/ml-k-means-algorithm/ [Pristupljeno 18. 6. 2024].